Usability Evaluation of Dialogue Design Strategies for Information Delivery

USABILITY EVALUATION OF DIALOGUE DESIGN STRATEGIES FOR INFORMATION DELIVERY

Nancie Davidson
CCIR, University of Edinburgh
The King's Buildings
Edinburgh
Nancie.Davidson@ccir.ed.ac.uk

Mervyn A. Jack
CCIR, University of Edinburgh
The King's Buildings
Edinburgh
Mervyn.Jack@ccir.ed.ac.uk

ABSTRACT

This paper describes a usability experiment designed to assess two strategies for information delivery in the context of a speech-enabled automated telephone service. Two versions of a service designed to provide promotional information to members of a frequent flyer programme were assessed. In the "List" version the full list of offers was played as a sequence of interruptible segments with "skip and search" navigation available through the use of meta commands. This was compared to the "Filter" version, in which users were asked to specify some or all of their journey requirements in order to filter out irrelevant information. The results indicate that the Filter approach is the more usable of the two. Participants preferred this version, rating it significantly higher in the usability questionnaire. Calls to the Filter version were also significantly shorter. Interestingly however, task performance was the same for both versions.

Keywords

Information delivery; Dialogue design; Usability; Speech.

1. INTRODUCTION

Recent advances in speech recognition technology have made the use of spoken natural language dialogues in automated telephone services an increasingly viable option. Typical applications include voice access to large databases of information, such as automatic directory assistance [1] or rail timetable information [6].

Within such services, where the purpose of the interaction is to access information, the question of how to present the results in a usable way arises. Intuitively, the information delivered to the user should be as concise as possible whilst satisfying their enquiry. In practice, where significant amounts of information are involved - for example lists of possible trains - this can present a considerable challenge in terms of usability.

If too much information is given, this may be tedious for the user and may make it difficult for them to extract the important part. If, on the other hand, not enough information is provided the user may be dissatisfied and have to ask for additional information, resulting in a longer interaction overall [6].

This paper describes an experiment carried out to evaluate the relative usability of two methods for information delivery within an automated telephone service. Two versions of a service designed to provide promotional information to members of a frequent flyer programme were assessed. In the "List" version the full list of offers was played as a sequence of interruptible segments with "skip and search" navigation available through the use of meta commands. This was compared to the "Filter" version, in which users were asked to specify some or all of their journey requirements in order to filter out irrelevant offers prior to presentation of the list.

2. DIALOGUE DESIGN

The context for the research was a spoken language dialogue service (SLDS) aimed at members of a frequent flyer programme. The purpose of this service was twofold; firstly, to provide members with basic account information (the number of points held, and a list of recent transactions), and secondly to encourage them to redeem their points, by providing access to promotional information.

Members of the frequent flyer programme targeted in the research earn points for every flight they purchase, which can then be used in exchange for free or discounted flights. Special offers are typically available on a large number of flights at any one time. In the sample used in the experiment a total of thirty European routes were available in addition to all U.K. domestic routes.

Figure 1 shows a top-level view of the service's dialogue architecture. This was common to both versions.

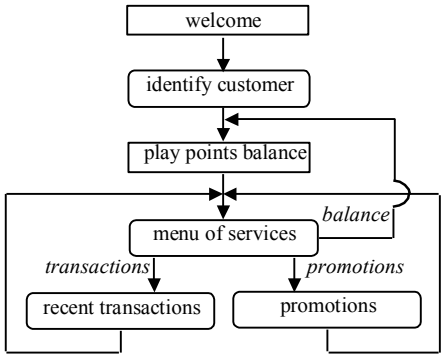

Figure 1. Dialogue Flow Chart

The two versions differed only in the promotions section of the service. In what was termed the "List" version, the full list of flight offers was played in summary form. In the "Filter" version callers were given the option to specify some or all of their journey requirements. Callers were first asked to specify a departure airport, or state that they "don't mind" where they fly from. This was then repeated for their destination, and in cases where either answer was not recognised confidently, this was followed by a confirmation stage. The information obtained was then used to "filter" out irrelevant offers from the list. Moreover, offers for which the caller did not have enough points were also excluded in this version.

In both versions, the final list of routes on offer was grouped into categories based on the number of points and cash required. The resulting four categories, each containing multiple routes, were then listed in order of ascending number of points and cash required. Navigation between categories was available via barge-in at any time using the meta commands "repeat", "previous", "next" or, to exit the list completely, "main menu". Callers were informed of this "skip and search" facility at the start of the listed information. Each set of routes was then introduced with a message indicating the total number of categories, together with details of the current category, as in for example "*Offer two of four. For £65 and 100 destination points you can fly on one of the following routes: Heathrow to Amsterdam, Brussels, Paris or Dublin; Edinburgh to Brussels…"*

The potential advantage of the List version is that it involves fewer dialogue stages, and allows callers to hear the full range of offers. However, the list of offers is long, and the onus is on the user to control the output and "pick out" offers that are relevant to them (both in terms of route and the number of points required). In contrast, the Filter version involves a more complex dialogue, but has a cognitively simpler end result.

3. METHODOLOGY

ISO 1998 [4] defines usability as *"the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use".* The ISO standard also notes that effectiveness and efficiency are often referred to as performance measures. Since satisfaction is subjective, this indicates that to establish usability the simultaneous measurement of both aspects is required. This is the premise on which the methodology (described more fully in [2]) is based.

The methodology employs an experimental approach, based on established techniques from experimental psychology. This is complemented by an emphasis on achieving as much realism within the experimental setting as is possible. Services under test are presented within the context of a realistic scenario, in which participants are encouraged to imagine themselves. Participants are then asked to undertake one or more tasks with a fully functional prototype that are typical of a real-life situation.

In studies with multiple design variants a repeated-measures approach is employed. Following each experience of a service, participants complete a usability questionnaire. This provides a quantitative measure of participants' attitude and is described in more detail in the following section.

3.1 Key Measures

3.1.1 Mean Attitude Score

This is derived from the usability questionnaire that participants are asked to complete after each telephone call. This questionnaire is a tool for assessing users' attitudes towards automated telephone services that has been developed and refined over a number of such experiments [3][2]. It consists of a set of proposal statements, each with a set of tick-boxes along a seven-point Likert scale [5], ranging from "strongly agree" through neutral to "strongly disagree". Statements in the questionnaire are balanced, positive and negative, to counteract the problem of response acquiescence set - the general tendency for respondents to agree with the statement offered. Once the polarity of the results is normalised, a measure of the mean attitude to the service can be obtained by averaging all the questionnaire results for participants who experienced that service.

In addition, the mean scores for individual statements can also be examined to highlight any aspects of the dialogue design which were particularly successful or which require improvement.

3.1.2 Explicit Preference

The second key measure is participants' *explicit preference* between the different versions of the service. This is obtained as part of a de-briefing interview at the end of the experiment.

3.1.3 Task Completion

The third key measure is the level of *task completion*: the proportion of participants who successfully accomplish each task. Recognition accuracy plays an important part in this, however task completion is also sensitive to other factors

such as the system's ability to elicit valid responses from the user, and to handle successfully any errors that occur. As such, it is an important objective measure of the effectiveness of the dialogue as a whole.

4. EXPERIMENT PROCEDURE

Participants made one telephone call to each version of the service, completing a usability questionnaire after each. This was followed at the end of the experiment by a de-briefing interview.

In each call participants were allocated the same persona and undertook three tasks. The first two tasks, the Balance and Recent Transactions tasks, were fixed across all participants. The third task, the Points Promotion task in which participants were asked to find out some information relating to the promotion, was varied across the group in order to reflect the various scenarios possible in real life.

Callers interested in flight promotions fall into four categories. There are those with no specific requirements, those with an exact route in mind, and those with *either* a preferred departure airport *or* a preferred destination. Each of these constraint types was represented in the experiment. The number of points held by participants was also varied during the experiment. Participants were allocated one of three possible values: not enough points for any of the offers, enough points for some of the offers, and enough points for all of the offers. All possible combinations of the experimental factors were balanced across the participant group.

Participants were asked to note down an answer for each task in addition to the automatic call logging provided by the system. This was used to determine whether users were able to extract the appropriate information from the List version. A total of 113 participants completed the experiment, in a group that was approximately balanced for age and gender. Three participant age groups were used (18-35 years, 36-49 years and 50+ years), and all participants were recruited from the general public.

5. RESULTS

5.1 Mean Attitude Score

Participants rated the usability of both versions above neutral (4 on the 7-point scale). The Filter version however, was rated *significantly* higher than the List version (mean for List = 4.35; mean for Filter = 4.56; ANOVA p=0.002). Participants found the Filter version more *efficient* (p=0.001), more *easy to use* (p=0.034) and were happier to *use it again* (p=0.020) compared to the List version. They were also significantly more positive in response to the statement *"The service was too fast for me"* (p=0.009) after using the Filter version.

The experimental variables *constraint type* and *number of points* were not found to have a significant effect.

5.2 Task Completion

Figure 2 shows the proportion of participants who successfully completed each task, for each version.

Task	List (% participants)	Filter (% participants)
Balance	96.9	94.7
Transactions	88.5	88.5
Promotions	75.2	79.6

Figure 2. Task Completion

There were no significant differences in the task completion figures for the two versions, for any of the tasks (McNemar). However, the *reasons* for failure in the promotions task were version-dependent All of those who failed using the List version (20 participants) successfully accessed the promotions information but failed to interpret it correctly and ticked the wrong answer on the task sheet as a result. Failures of this type were less frequent in the Filter version but due to the additional dialogue stages there was a higher incidence of dialogue failure resulting in "breakout" to an agent. Breakouts were largely due to a mixture of false rejections by the recogniser and out-of-grammar utterances on the part of the user.

5.3 Explicit Preference

Participants who noticed a difference between the versions (74.3%) were asked which version they preferred. In total, 53.1% of participants selected the Filter version as their preferred version, whilst 12.4% chose the List version. A chi-square test confirmed that this distribution of responses was very unlikely to occur by chance ($p<0.001$). Reasons given for preferring the Filter version were mainly that it was quicker and/or easier, and that it did not involve dealing with long lists of information.

5.4 Other Results

5.4.1 Use of Navigation Meta Commands

Figure 3 shows the percentage of participants who used each navigation meta command at least once during the promotions section of the dialogue.

Command	List (% participants)	Filter (% participants)
Repeat	6.2	0.0
Previous	3.5	0.0
Next	18.6	0.9
Main Menu	9.7	0.9

Figure 3. Use of Navigation Meta Commands

Use of these commands, although higher in the List version as might be expected, was low in both versions. Usage was spread across the various experimental groups, with no discernible pattern.

5.4.2 Call Duration

Calls to the Filter version were on average significantly shorter than calls to the List version ($p<0.001$). The average

call duration for the Filter version was 187 seconds, compared to 230 seconds for the List version.

The experimental variables *constraint type* and *number of points* did not have a significant effect on call duration. This was a little surprising given that, for example, in the case where participants did not have enough points to qualify for any of the offers, the Filter version played a single message to that effect, whilst the List version played the full set of offers. However, closer examination showed that whilst calls to the List version were longer across most of the experimental conditions, other factors, such as participants' behaviour in the other tasks (opting for example to listen to the list of recent transactions more than once) obscured any potential relative difference in call duration resulting from the different experimental conditions. Information on the duration of individual tasks was unfortunately not available.

6. DISCUSSION

Although there was some use of the "skip and search" navigation commands in the List version, participants on the whole did not make use of this facility, preferring to remain passive and listen to the full list of offers. Of course, the degree of benefit involved in using these commands depended on the number of points held and the level of constraint in the enquiry. However, there was no evidence to suggest that callers' individual circumstances influenced their use of these commands. It may have been that the structure of the listed information was not as transparent as hoped, with the result that participants did not feel confident enough to "skip" offers, opting instead to listen to the complete list.

As a result of this, call duration was significantly shorter in the Filter version. In terms of dialogue efficiency the cost of additional stages was outweighed on average by the benefit of not playing the full list of offers in every call. Interestingly however, task completion was similar in both versions; in this case the advantage of a reduced cognitive load on the user was not enough to outweigh the cost of additional dialogue stages.

Participants rated the Filter approach to information delivery significantly higher than the List approach. Significantly more participants also selected the former when asked which version they preferred.

Again however, neither the number of points held, nor the level of constraint in the enquiry had a significant effect on the results. It might have been expected, for example, that the preference for the Filter version would be less pronounced amongst participants with enough points for all the offers compared to those in other groups, since under these circumstances the amount of information played out by both versions was the same (and moreover the Filter version involved extra dialogue stages). This, however, was not the case. Similarly, there was no evidence that a more specific set of journey requirements resulted in a greater preference for the Filter version.

This is an interesting result. One possible interpretation is that users like to be asked for their preferences, regardless of the degree to which this affects the volume of output. However, overall it is not clear whether participants' preference for the Filter version is attributable to this feature, or to the reduction in call time compared to the List version. It would be interesting to investigate a third alternative, in which the information delivered was filtered based *only* on the number of points. This would have potential efficiency advantages compared to the List version, but without the cost of additional dialogue stages. However, it would also remove participants' ability to state their preferences. Further work is required to investigate the issues arising from this experiment in more detail.

7. ACKNOWLEDGEMENTS

This work was carried out as part of EC IST Project Spotlight. The authors wish to acknowledge Nuance Communications for use of their recognition software, and bmi for providing the context for the research.

8. REFERENCES

[1] Boves, L., Jouvet, D., Sienel, J., de Mori, R., Béchet, F., Fissore, L. and Laface, P. 2000. ASR for automatic directory assistance: the SMADA project. In Proc. ASR 2000 ISCA Tutorial and Research Workshop on Automatic Speech Recognition, September, Paris.

[2] Davidson, N., McInnes, F.R. and Jack, M.A. (in press). Usability of dialogue design strategies for automated surname capture. Speech Communication.

[3] Dutton, R., Foster, J., Jack, M.A. and Stentiford, F. 1993. Identifying usability attributes of automated telephone services. In Proc.EUROSPEECH'93, 1335-1338.

[4] ISO (International Standardisation Organisation). 1998. ISO-9241: Ergonomic requirements for office work with visual display terminals (VDTs). Part 11: Guidance on usability. http://www.iso.org

[5] Likert, R. 1932. A technique for the measurement of attitudes. Archives of Psychology, 140.

[6] Lamel, L., Rosset, S., Gauvain, J., Bennacef, S., Garnier-Rizet, M. and Prouts, B. 2000. The LIMSI ARISE system. Speech Communication, 31, 339-353.